# Visual Workflow Recognition Using a Variational Bayesian Treatment of Multistream Fused Hidden Markov Models

Sotirios P. Chatzis and Dimitrios Kosmopoulos

***Abstract*—In this work, we provide a variational Bayesian (VB) treatment of multistream fused hidden Markov models (MFHMMs), and we apply it in the context of active learning-based visual workflow recognition. Contrary to training methods yielding point estimates, such as maximum likelihood or maximum a posteriori training, the VB approach provides an estimate of the posterior distribution over the MFHMM parameters. As a result, our approach provides an elegant solution towards the amelioration of the overfitting issues of point estimate-based methods. Additionally, it provides a measure of confidence in the accuracy of the learned model, thus allowing for the easy and cost-effective utilization of active learning in the context of MFHMMs. Two alternative active learning algorithms are considered in this paper: query by committee, which selects unlabeled data that minimize the classification variance, and a maximum information gain method which aims to maximize the alteration in model variance by proper data labeling. We demonstrate the efficacy of the proposed treatment of MFHMMs by examining two challenging workflow recognition scenarios, and we show that the application of active learning, which is facilitated by our VB approach, allows for a significant reduction of the MFHMM training costs.**

[1]

## I. Introduction

Human behavior understanding in video sequences is a research field rapidly gaining momentum over the last few years. This is mainly due to its fundamental applications in automated video indexing, virtual reality, human-computer interaction and smart monitoring. Especially, throughout the last years we have seen an increasing need for assisting and extending the capabilities of human operators in remotely monitored large and complex spaces such as public areas, airports, railway stations, parking lots, bridges, tunnels, etc. The last generation of surveillance systems was designed to utilize multiple video streams from heterogeneous sensors to automatically assess the ongoing activities in large monitored environments, flagging and presenting to the operator suspicious events as they happen in order to prevent dangerous situations [1], [2].

In this work, we focus on *visual workflow recognition*; workflows are comparatively structured processes, in contrast to monitoring stations or airports, and it is more realistic to believe that workflows can be modeled using computer vision and machine learning. The identified deviations from a predefined workflow possibly indicate security and safety related events and will be automatically highlighted. Distributed smart workflow monitoring is applicable to mass production or large-scale enterprises like industrial plants which have a clear need for automated supervision services to guarantee safety, security, and quality by enforcing adherence to predefined procedures for production or services. Such supervision services are frequently of vital importance for the enterprise because, apart from cost reduction, timely detection of safety and security concerns may prevent injuries and even fatalities.

The complexity of detection and tracking of moving objects under occlusions in a typical structured environment requires more than a single camera and features that will not result from an error-prone tracker. Multiple cameras provide a wider coverage of the scene and redundant data that help solve occlusions and improve accuracy. Furthermore, the high diversity and complexity of the behaviors which need to be monitored requires new learning methods that will be able to fuse information from multiple streams. Finally, the limited availability in model training data, due to the prohibitively high costs of capturing and annotating behavioral data from real (e.g., industrial) installations, necessitates utilization of an active learning framework, allowing for the exploitation of unlabeled data to improve the classification performance of the trained models.

Hidden Markov models (Fig. 1a) are an extremely popular means of modeling a stream of sequential data, and are vastly adopted in behavioral analysis applications [2]. Using information from multiple streams of data pertaining to the same sequence of events has been shown to allow for a significant performance enhancement of HMM-based event analysis and detection models [3]–[6]. Modern multimedia capturing and processing technologies have rendered insignificant the main hurdle of the additional computational requirements imposed by multisensor systems. [2]. However, the reliability of the sensors is never explicitly considered. Hence, in a video surveillance system that employs multiple sensors, the problem of selecting the most appropriate sensor or set of sensors to perform a certain task often arises. Consequently, the first and most straightforward solution of *early integration* [7], which consists in merging all the observations related to all the streams into one large stream (frame by frame), and modeling it using a single HMM, is less than satisfactory (Fig. 1b). To resolve this problem, an adaptive multicue multicamera information fusion framework based on *democratic integration* [8] is presented in [9]. Fusion is performed by taking into

account sensor reliability, yet there is no direct sensor quality assessment. Instead, the reliability of a source is estimated by measuring the distance between each source estimate and the fused estimate, which is determined by the sources estimates. This is based on the assumption that the majority of sensors are producing reliable estimates, which cannot always be taken for granted.

A different probabilistic framework for multistream data fusion is the multistream HMM approach [10], under which each stream is modeled separately using its own HMM. Then, analysis of the observed data can be conducted by creating a special HMM, recombining all the single stream HMM likelihoods at various specific temporal points. Obviously, depending on the specific selection of these recombination points, different solutions arise. For instance, in *coupled hidden Markov models (CHMMs)* [5] (Fig. 1e), two component HMMs are linked by the dependence of their hidden states. However, in many applications where the component HMMs do not consist of many states, such as in cases of audio-visual data, this dependence assumption is not strong enough to capture the statistical correlations between the multiple streams. In the state-synchronous multistream HMM (Fig. 1c) the streams are assumed to be synchronized. Each stream is modelled using an individual HMM; the postulated streamwise HMMs share the same state dynamics. As a result, this approach provides only limited sequential data modeling flexibility.

In [11], a *parallel hidden Markov model (PaHMM)* has been proposed (Fig. 1d), which factorizes the state space into multiple independent temporal processes without causal connections in-between. Nevertheless, the assumption of the different temporal processes being independent of each other is clearly invalid in most cases, especially when dealing with group or interactive activities.

*Multistream fused HMMs (MFHMMs)* is another promising method for multistream data modeling [12] (Fig. 1f). Like coupled HMMs and mixed-memory HMMs, an MFHMM consists of multiple HMMs. However, unlike the previous methods, the connections between the component HMMs are chosen based on a probabilistic fusion model, which is optimal according to the maximum entropy principle and a maximum mutual information criterion for selecting dimensionality reduction transforms. As a consequence, the MFHMM has several desirable features: a) It has simpler and faster training and inference algorithms than the previous models; b) if one of the component HMMs fails due to noise or a probable malfunction of the sensor capturing the related observations stream, the rest of the constituent HMMs can still work properly; and c) it still retains the crucial information about the interdependencies between the multiple data streams, which coupled HMMs tend to neglect.

In this paper, motivated by the aforementioned advantages of MFHMMs, we consider their application to the addressed problem of visual behavioral analysis and monitoring from multiple visual inputs in structured environments *(workflow recognition)*. In the existing literature, MFHMM is treated under a maximum-likelihood (ML) framework, using the expectation-maximization (EM) algorithm (see, e.g., [12]). Even though maximum-likelihood is a common, and, in general, reliable approach for estimation of probabilistic generative models, it suffers from the undesirable property of being ill-posed since the likelihood function is unbounded from above [13]–[15]. This fact might result in several very significant deficiencies, especially in cases of limited training data availability; this is quite the case regarding the applications we focus on in this paper, as training data from real installations are difficult to collect, and very expensive to process and annotate. As a result, using ML to train a set of generative models for behavioral analysis and detection in such a context might result in an unstable training procedure, yielding poor model estimates, with high overfitting proneness; it could even lead to yielding infinities in the likelihood function, associated with the collapsing of the bell-shaped component distributions onto individual data points, and, hence, resulting in singular or near-singular covariance matrices [15].

To address these issues, in this work we introduce a Bayesian treatment of MFHMMs, overcoming the problems of ML approaches elegantly, by marginalizing over the model parameters with respect to appropriate priors, and maximizing the resulting marginal likelihood of the model to obtain the optimal model size. Our approach is based on variational approximation methods [16], which have recently emerged as a deterministic alternative to Markov chain Monte-Carlo (MCMC) algorithms for doing Bayesian inference for probabilistic generative models [17], [18], with better scalability in terms of computational cost [19]. Variational Bayesian (VB) inference has been previously applied to a number of probabilistic inference models, including relevance vector machines [20], autoregressive models [21], [22], mixtures of Gaussians and Student's-$t$ distributions [23], [24], mixtures of factor analyzers [25], [26], and HMMs [27], [28], thereby ameliorating the singularity and overfitting problems of ML approaches in an elegant and computationally efficient manner.

Since variational Bayes provides a full posterior distribution over the treated model parameters, the proposed approach allows for the extraction of a reliable measure of confidence in the obtained estimates of a trained MFHMM. This is yet another significant advantage of the proposed VB treatment of MFHMMs, as it allows for the easy and computationally efficient introduction of the MFHMM in the context of an elegant active learning framework. Indeed, as we shall discuss in the following sections of this paper, under the proposed variational Bayesian regard, well-known active learning criteria can be easily implemented for MFHMMs, while previously they were either computationally inefficient or intractable (when considering point-estimated MFHMMs) [29], [30]. Therefore, the introduction of the VB machinery does also allow for the exploitation of effective active learning methodologies so as to significantly reduce the training costs of MFHMMs, by efficiently utilizing pools of cheap to acquire unlabeled data.

The remainder of this paper is organized as follows: In Section II, the proposed variational Bayesian treatment of MFHMMs is introduced, and the related model inference and prediction algorithms are derived. In Section III, the proposed approach is examined in the context of the active learning framework. As we show, the proposed VB treatment of MFHMMs allows for the efficient utilization of effective

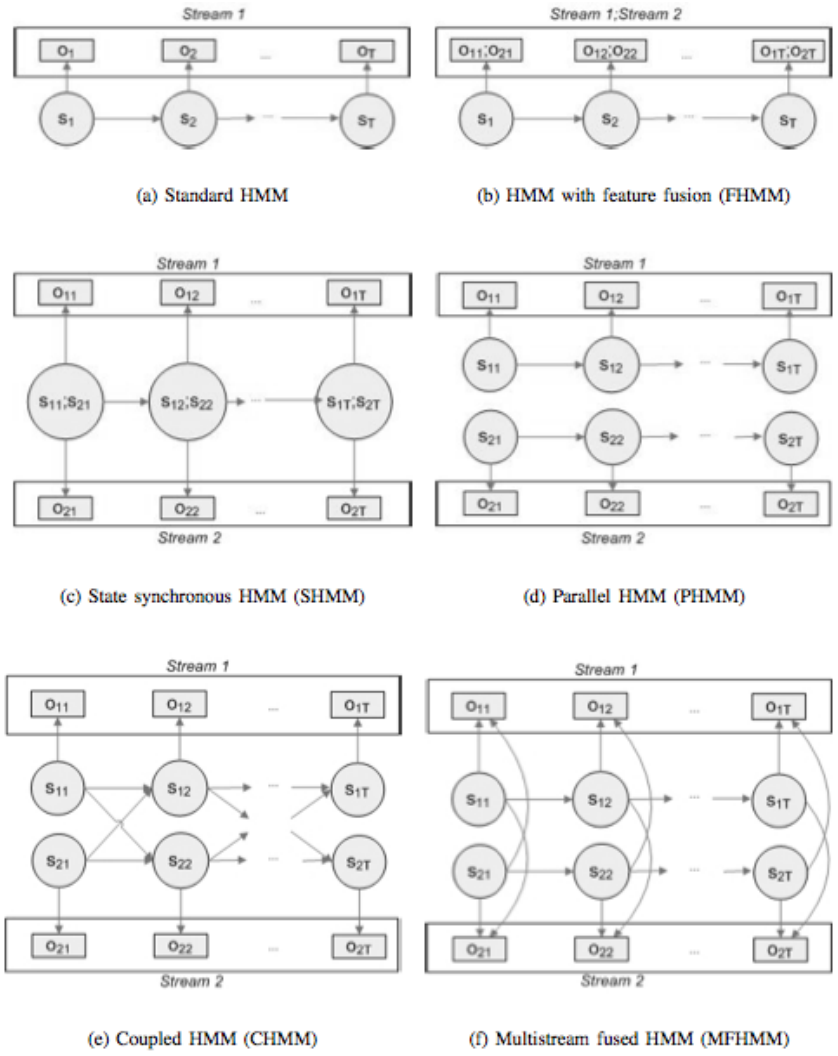



Figure 1. Various fusion schemes using the HMM framework for two streams. The $s$, $o$ stand for the states and the observations respectively. The first index marks the stream and the second the time.

active learning algorithms, which would be either impossible or computationally burdensome to apply when considering point-estimated MFHMMs. In Section IV, we examine the efficacy of the proposed approach considering two challenging visual workflow recognition scenarios using publicly available datasets. Finally, in the concluding section of this paper, we summarize and discuss our results.

## II. A Variational Bayesian Approach Towards MFHMMs

### A. Multistream Fused Hidden Markov Models

Consider $M$ tightly interdependent time series (streams), $X = \{X^m\}_{m=1}^M$, with $X^m = \{\boldsymbol{x}_t^m\}_{t=1}^T$. Assume that the constituent streams $X^m$, $m \in \{1, \ldots, M\}$ of $X$, can be modeled by $M$ independent (*streamwise*) HMMs, with their corresponding hidden state sequences denoted as $S^m = \{s_t^m\}_{t=1}^T$. Then, we have

$$p(X^m) = \sum_{S^m} p(X^m, S^m) \tag{1}$$

where

$$p(X^m, S^m) = p(s_1^m) p(\boldsymbol{x}_1^m | s_1^m) \prod_{t=2}^{T} p(s_t^m | s_{t-1}^m) p(\boldsymbol{x}_t^m | s_t^m) \quad \forall m \tag{2}$$

and $p(\boldsymbol{x}_t^m | s_t^m)$ are the state-conditional likelihoods of the models, usually selected to be mixtures of Gaussians, or mixtures of Student's-$t$ densities [31]. In the following, we shall be denoting as $\boldsymbol{\pi}^m = (\pi_n^m)_n$ the initial state probabilities vector of the $m$th postulated *streamwise* HMM, with

$$\pi_n^m \triangleq p(s_1^m = n)$$

and as $\boldsymbol{A}^m = (a_{ij}^m)_{i,j}$ the corresponding state transition probabilities matrix, with

$$a_{ij}^m \triangleq p(s_t^m = j | s_{t-1}^m = i) \; \forall t$$

The problem addressed by MFHMMs is how to construct a new structure linking the postulated *streamwise* HMMs together, that will be giving an optimal approximation of the joint probability of the stream data, $p(X) = p\left(\{X^m\}_{m=1}^M\right)$

[12]. For this purpose, MFHMMs take advantage of the fact that the streams $\{X^m\}_{m=1}^M$ can be separately modeled by individual HMMs. Then, to capture the statistical dependence between these streams, a set of transforms $\boldsymbol{w}^m \triangleq g(X^m)$ is introduced, such that the joint probability $p(\{\boldsymbol{w}^m\}_{m=1}^M)$ can be more easily calculated compared to $p(\{X^m\}_{m=1}^M)$. On the basis of this regard, the MFHMM obtains an optimal approximation of $p(X)$ according to the *maximum entropy principle*, given by [32]

$$p(X) \approx \tilde{p}(X) \tag{3}$$

where

$$\tilde{p}(X) = \tilde{p}\left(\{X^m\}_{m=1}^M\right) \triangleq \frac{p\left(\{\boldsymbol{w}^m\}_{m=1}^M\right)}{\prod_{m=1}^M p(\boldsymbol{w}^m)} \prod_{m=1}^M p(X^m) \tag{4}$$

Selection of a proper expression for the transforms $\boldsymbol{w}^m$ is conducted on the basis of the *maximum mutual information* (MMI) criterion [33], a criterion which has been also used for discriminative training of HMMs with quite a success [34]. MMI criterion essentially comprises minimization of the Kullback-Leibler divergence $\mathrm{KL}(p||\tilde{p})$ between the exact distribution $p(X)$ and the approximate distribution $\tilde{p}(X)$, where

$$\begin{aligned}\mathrm{KL}(p||\tilde{p}) = -\int \mathrm{d}X^1 \ldots \int \mathrm{d}X^M \\ \times p\left(\{X^m\}_{m=1}^M\right) \log \frac{\tilde{p}\left(\{X^m\}_{m=1}^M\right)}{p\left(\{X^m\}_{m=1}^M\right)}\end{aligned} \tag{5}$$

It can be shown [12], that by application of the MMI criterion, and considering that all the fused data streams of the MFHMM are (a priori) of equal reliability, Eq. (4) yields

$$\tilde{p}(X) = \tilde{p}\left(\{X^m\}_{m=1}^M\right) = \frac{1}{M} \sum_{m=1}^M p(X^m) \prod_{r \neq m} p(X^r|\hat{S}^m) \tag{6}$$

In Eq. (6), $\hat{S}^m$ are the state sequence estimates of the available stream data, obtained by application of the Viterbi algorithm [15] on the individual *streamwise* HMMs comprising the postulated MFHMM. Regarding the *coupling densities* $p(X^r|\hat{S}^m)$, from the conditional independence property of the Markovian chain, we yield

$$p(X^r|\hat{S}^m) = \prod_{t=1}^T p(\boldsymbol{x}_t^r|\hat{s}_t^m) \tag{7}$$

The probabilities $p(\boldsymbol{x}_t^r|\hat{s}_t^m)$ of the MFHMM can be modeled by means of mixtures of Gaussian or Student's-$t$ densities, similar to the state-conditional likelihoods of the *streamwise* HMMs. Note that for each possible value, say $i$, of $\hat{s}_t^m$, a different *coupling density model* $p(\boldsymbol{x}_t^r|\hat{s}_t^m = i)$ is to be postulated. Hence, if we consider $N$-state *streamwise* HMMs, there is a total of $N$ different finite mixture models that must be trained to model the *coupling densities* $p(\boldsymbol{x}_t^r|\hat{s}_t^m),\ \forall r, m$.

### B. Variational Bayesian Inference for the MFHMM

Bayesian treatments of probabilistic generative models comprise introduction of a set of prior distributions over the model parameters and further maximization of the model's log marginal likelihood (log evidence). For convenience, usually conjugate priors are preferred, as this selection greatly simplifies inference and interpretability [16]. However, due to the complexity of the MFHMM, exact Bayesian inference for our model is intractable. Nevertheless, the choice of conjugate exponential prior distributions for the model parameters allows for the derivation of an elegant variational framework.

Let us consider a model $p(X|\Psi)$ treated under the variational Bayesian paradigm. Let $p(\Psi)$ be the conjugate prior imposed on the model, and $X$ be the used set of training data. Variational Bayesian inference is conducted by introducing an *approximate* (variational) posterior over the model parameters $q(\Psi)$, and considering the well-known equality for the log evidence, $\log p(X)$ [19]

$$\log p(X) = F(q) + \mathrm{KL}(q||p) \tag{8}$$

where

$$F(q) = \int \mathrm{d}\Psi q(\Psi) \log \frac{p(X, \Psi)}{q(\Psi)} \tag{9}$$

Since the KL divergence term in (8) is a non-negative quantity, $F(q)$ comprises a strict lower bound of the log evidence, i.e.

$$\log p(X) \geq F(q) \tag{10}$$

and would become exact if $q(\Psi) = p(\Psi|X)$. Hence, maximizing the lower bound of the log evidence (*variational free energy*), $F(q)$, so that it becomes as tight as possible, i.e. minimizing the KL divergence between the true and the variational posterior, a good variational inference scheme is obtained. In other words, variational Bayes can be summarized under the maximization scheme

$$q(\Psi) = \mathrm{argmax}_q F(q) \tag{11}$$

It is worthwhile to note that the variational posteriors obtained by optimization of the variational free energy $F(q)$ are only an approximation of the actual posterior densities $p(\Psi|X)$. However, the variational Bayesian approach allows for considerably better scalability in terms of computational cost [19] compared to exact Bayesian inference using Markov chain Monte-Carlo (MCMC) algorithms, which becomes of practical importance in applications requiring fast processing of high-dimensional large-scale datasets.

As the MFHMM consists of two fundamental "building blocks," the *streamwise* HMMs, and the *coupling models*, variational Bayesian inference for this model can be degenerated into two separate procedures: (a) variational Bayes for the postulated *(streamwise)* HMMs; and (b) variational Bayes for the postulated finite mixture models *(coupling models)*. Below, we provide an outline of the proposed variational Bayesian treatment of the MFHMM.

### C. Variational Posteriors

Consider an MFHMM modeling $M$ tightly interdependent time series, $\{X^m\}_{m=1}^M$. For simplicity, and without any loss of generality, we assume that all the observed time series have the same length, $T$, i.e. $X^m = \{\boldsymbol{x}_t^m\}_{t=1}^T$. The variational posteriors of the postulated MFHMM can be derived as follows.

*1) Streamwise HMM training:* Initially, $M$ individual HMMs are trained independently (one for each stream), by means of the VB algorithm, as described, e.g., in [27]. These are the constituent *streamwise* HMMs of our model, with obtained variational posteriors $q(\Psi^m)$, where $\Psi^m$ are the parameters of the $m$th constituent *streamwise* HMM, $m = 1, \ldots, M$. For simplicity, and without any loss of generality, we consider $N$-state *streamwise* HMMs.

Specifically, VB inference for the *streamwise* HMMs of our model is conducted by imposing Dirichlet priors over the initial state and state transition probabilities of the models:

$$p(\boldsymbol{\pi}^m) = \mathcal{D}(\pi_1^m, \ldots \pi_N^m | \phi_1^m, \ldots, \phi_N^m) \tag{12}$$

$$p(\boldsymbol{A}^m) = \prod_{i=1}^{N} \mathcal{D}(a_{i1}^m, \ldots, a_{iN}^m | \upsilon_{i1}^m, \ldots, \upsilon_{iN}^m) \tag{13}$$

The observation emission probabilities of the hidden states of the models are taken as finite mixtures of Gaussian or Student's-$t$ distributions. Considering for simplicity $K$-component mixture models, we impose a Dirichlet prior over their mixture component weights, of the form:

$$p(\boldsymbol{\Delta}^m) = \prod_{i=1}^{N} \mathcal{D}(\delta_{i1}^m, \ldots, \delta_{iK}^m | \epsilon_{i1}^m, \ldots, \epsilon_{iK}^m) \tag{14}$$

and a joint Normal-Wishart prior over the means and precision matrices of the (Gaussian or Student's-$t$) mixture component densities:

$$p(\{\boldsymbol{\mu}_{ij}^m, \boldsymbol{R}_{ij}^m\}_{i,j=1}^{N,K}) = \prod_{i=1}^{N} \prod_{j=1}^{K} \mathcal{NW}(\boldsymbol{\mu}_{ij}^m, \boldsymbol{R}_{ij}^m | \lambda_{ij}^m, \boldsymbol{\gamma}_{ij}^m, \eta_{ij}^m, \boldsymbol{Q}_{ij}^m) \tag{15}$$

As a result of choosing to impose conjugate priors over the parameters of our model, the resulting variational posteriors of the model parameters take the same functional form as their corresponding priors [19]. Complete derivations of these posteriors have been provided in one of our previous works [27], and hence we refrain from repeating them here for brevity.

*2) Sequence decoding:* The best hidden state sequences $\hat{S}^m$ of the *streamwise* HMMs, corresponding to the used training data $X^m$, are found using the *VB Viterbi algorithm* [27]. The VB Viterbi algorithm comprises maximization of the approximate (variational) posterior expectation of $\log p(X^m, S^m | \Psi^m)$:

$$\hat{S}^m = \operatorname{argmax}_{S^m} \int \mathrm{d}\Psi^m q(\Psi^m) \log p(X^m, S^m | \Psi^m) \tag{16}$$

where $\log p(X^m, S^m | \Psi^m)$ is defined in (2) (for details, refer to [27]).

*3) Coupling models training:* Finally, the *coupling models* are obtained. This problem is equivalent to postulating one finite mixture model (with Gaussian or Student's-$t$ densities) for each of the distributions $p(\boldsymbol{x}_t^r | \hat{s}_t^m = i)$, $\forall i \in \{1, \ldots, N\}, r, m \in \{1, \ldots, M\}, r \neq m$, and subsequently employing variational Bayes to obtain the variational posteriors $q(\Psi_i^{r,m})$ over their parameters sets $\Psi_i^{r,m}$. The complete derivations of the VB training algorithm for finite mixtures of Gaussian densities can be found in [16], while for the case of Student's-$t$ densities they are provided in [24].

### D. Hidden State Sequence Estimation Algorithm

Essentially, this is the problem of maximizing

$$\{\hat{S}^m\}_{m=1}^M = \operatorname{argmax}_{\{S^m\}_{m=1}^M} \int \mathrm{d}\Psi q(\Psi) \times \log p(\{X^m, S^m\}_{m=1}^M | \Psi) \tag{17}$$

where $\Psi \triangleq \{\{\Psi_i^{r,m}\}_{i=1}^N, \Psi^m\}_{m,r=1,m\neq r}^M$. Then, following the related results of [12], and assuming that all the postulated *streamwise* HMMs are of the same reliability, using (3) and (6) we have that (17) eventually reads

$$\hat{S}^m = \operatorname{argmax}_{S^m} \int \mathrm{d}\Psi^m q(\Psi^m) \times \log \left[ p(X^m, S^m | \Psi^m) \prod_{r \neq m} p(X^r | S^m; \Psi_{S_m}^{r,m}) \right] \tag{18}$$

Comparing the result (18) with (16), we directly observe that estimation of the optimal state sequences $\hat{S}^m$ for the MFHMM effectively boils down to merely an application of the VB Viterbi algorithm, with the probabilities

$$p(X^m | S^m) = \prod_{t=1}^{T} p(\boldsymbol{x}_t^m | \boldsymbol{s}_t^m) \; \forall m \tag{19}$$

of the single-HMM Viterbi algorithm being now replaced with the products

$$\prod_{\forall r} p(X^r | S^m) = \prod_{\forall r} \prod_{t=1}^{T} p(\boldsymbol{x}_t^r | \boldsymbol{s}_t^m) \; \forall m \tag{20}$$

in which expression the quantities $p(X^m | S^m)$ are given by the postulated *streamwise* HMMs, and the quantities $p(X^r | S^m)$, $r \neq m$, are given by the *coupling* models.

### E. Predictive Probability

The ultimate goal of Bayesian learning is, given a set of test data, to perform density estimation with respect to the learned model. Let us suppose the test data $Y = \{Y^m\}_{m=1}^M$, with $Y^m = \{\boldsymbol{y}_t^m\}_{t=1}^T$, and an MFHMM trained using the training data $X$, with obtained variational posterior $q(\Psi)$. The variational (approximate) predictive density of the given test data with respect to the considered MFHMM is given by

$$p(Y | X) = \int \mathrm{d}\Psi q(\Psi) p(Y | \Psi) \tag{21}$$

yielding

$$p(Y | X) = p\left(\{Y^m\}_{m=1}^M | X\right) \approx \frac{1}{M} \sum_{m=1}^{M} q(Y^m) \prod_{r \neq m} q(Y^r | \hat{S}^m) \tag{22}$$

where

$$q(Y^r | \hat{S}^m) = \prod_{t=1}^{T} q(\boldsymbol{y}_t^r | \hat{\boldsymbol{s}}_t{}^m) \tag{23}$$

while $q(Y^m)$ and $q(\boldsymbol{y}_t^r | \hat{\boldsymbol{s}}_t^m)$ are the *predictive densities* of the *streamwise* HMMs and the *coupling models*, respectively, comprising the trained MFHMM, which can be obtained based on the VB treatments of these models (see, e.g., the descriptions in [27] regarding the *streamwise* HMMs, and the discussions in [16] regarding the *coupling* models).

## III. How Does VB Facilitate MFHMM-Based Active Learning?

As we have already discussed, due to the prohibitively high costs of capturing and annotating behavioral data from real installations, measures have to be taken to avoid severe MFHMM training algorithm instabilities (e.g., yielding singular covariance estimates). Variational Bayes serves us well towards the achievement of this goal. However, another significant repercussion of the shortage in training data regards the high chances of the trained model manifesting a notably poor generalization performance [35].

To remedy this issue, we employ in this work the concept of active learning. Active learning is based on the notion that the performance of the learners (here, MFHMMs) might be considerably improved if the learners could actively participate in the learning process [36]. That is, contrary to conventional supervised learning, where the learner "passively" receives the labeled data and generates a learned model, we would like to introduce a framework for identifying a subset of a pool of unlabeled examples that would be most informative if the associated labels were available and incorporate them in the learning procedure. Hence, the proposed active learning methodology comprises two basic procedures: first, selection of the most informative samples from a pool of unlabeled data; and, second, labeling of these samples and introduction into the model training procedure of the MFHMM.

Under the proposed Bayesian treatment of the MFHMM, the informativeness of a new data point can be assessed analytically by viewing unlabeled sample selection as an information extraction process: we select the data that gives us maximum information about the pool of unlabeled samples; in other words, we apply an information gain criterion. Since variational Bayes yields a posterior over the model parameters $\Psi$, information gain after augmenting an unlabeled data into the training set can be expressed in the context of information theory: "How much information about $\Psi$ can be obtained if we add an unlabeled data $X^*$ into the training set?"

Indeed, let us consider $C$ modeled behavioral classes, each one represented by a postulated MFHMM. Following [37], we measure the information gain obtained by adding an unlabeled data $X^*$ into the training set by means of the KL divergence between the posterior density of the MFHMM parameters $\Psi$ obtained after augmenting the unlabeled data $X^*$ into the training set and before the augmentation [37], defined as

$$G(X^*) \triangleq \sum_{c^*=1}^{C} KL\Big(p(\Psi_{c^*}|X^*, X)||p(\Psi_{c^*}|X)\Big)p(c^*|X^*; X) \tag{24}$$

In (24), $p(\Psi_{c^*}|X^*, X)$ is the variational posterior of the $c^*$th postulated MFHMM (modeling the $c^*$th behavioral class), obtained after augmenting the unlabeled data $X^*$ into the training data of the class; $p(\Psi_{c^*}|X)$ is the variational posterior of the $c^*$th postulated MFHMM obtained before augmenting any unlabeled data; and, finally, $p(c^*|X^*; X)$ is the *a posteriori* probability of the $c^*$th class regarding the unlabeled sample $X^*$, which, considering all the classes of equal *a priori* probability, is given by

$$p(c^*|X^*; X) = \frac{p_{c^*}(X^*|X)}{\sum_{k=1}^{C} p_k(X^*|X)} \tag{25}$$

where $p_k(X^*|X)$ is the (variational) predictive probability of the data $X^*$ with respect to the $k$th class MFHMM, defined in (22).

In essence, $G(X^*)$ seeks labels that can most shrink or expand (i.e., change) the model variance; thus, the information gain obtained by this measure is defined in terms of the possible change in the model variance, which has been shown to be more appropriate than other related information gain metrics [38], as well as other candidate unlabeled data selection strategies, e.g., the query by committee (QBC) approach [39], for comparably low computational costs. Finally, in regards to the labeling decision for the unlabeled samples selected to be incorporated in the model training procedure, this can be simply effected by maximization of the *a posteriori* probabilities (25) of the selected data points over the class labels $c^*$.

In our experimental investigations, apart from the information gain criterion (24), we shall also consider the QBC approach as another alternative for the conduction of active learning in the context of the variational Bayesian MFHMM. In the framework of QBC [39], [40], the informativeness of an example is measured by computing the classification variance with respect to the entire space of possible models consistent with the training data thus far. Since this computation is practically infeasible, the QBC algorithm approximates the entire space by randomly sampling the posterior distribution of the model parameters obtained from model training. These randomly selected models serve as a "committee" of classifiers to classify each unlabeled example. Then, the classification variance is measured by computing the disagreement over the classifications obtained by the classifiers comprising the committee. The data samples with the strongest disagreement among the committee are selected for labeling.

In this work, this degree of disagreement shall be measured via the KL divergence, measuring the average distance of the class posterior density resulting from each committee member to their mean value. Specifically, let us denote as $\{\hat{\Psi}_{\xi}^{c}\}_{\xi=1}^{\Xi}$ a set of $\Xi$ instances of the trained MFHMM model of the $c$th class, with variational posterior $q(\Psi^c)$, obtained by sampling $q(\Psi^c)$ $\Xi$ consecutive times. Then, the score of an unlabeled data $X^*$ given by the sampled committee of experts is given by

$$score(X^*) = \frac{1}{\Xi}\sum_{\xi=1}^{\Xi} KL\left[p\left(c^*|X^*, \hat{\Psi}_{\xi}^{c^*}\right)||p_{avg}\left(c^*|X^*\right)\right] \tag{26}$$

where

$$p_{avg}\left(c^*|X^*\right) = \frac{1}{\Xi}\sum_{\xi=1}^{\Xi} p\left(c^*|X^*, \hat{\Psi}_{\xi}^{c^*}\right) \tag{27}$$

and, considering all the classes of equal *a priori* probability, we have

$$p\left(c|X^*, \hat{\Psi}_{\xi}^{c}\right) = \frac{p_c(X^*|\hat{\Psi}_{\xi}^{c})}{\sum_{k=1}^{C} p_k(X^*|\hat{\Psi}_{\xi}^{k})} \tag{28}$$

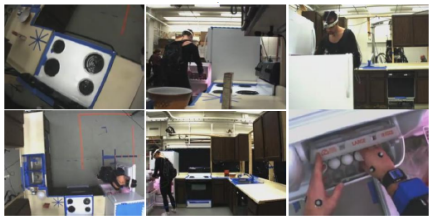


Figure 2. Different camera views in the CMU Multi-Modal Activity Database (from [41]). We used the cameras 7151020 (first in second row) and 7151062 (second in first row)

where $p_c(X^*|\hat{\Psi}_\xi^c)$ is the predictive probability of the MFHMM of the $c$th class with respect to $X^*$, and $c^*$ in (26) and (27) is the class the maximizes (28) for the given committee member $\xi$ and predictive point $X^*$.

## IV. Experimental Results

To experimentally verify the proposed approach, we have used some public benchmark datasets involving action recognition of humans, namely the CMU-MMAC and workflow recognition (WR) databases.

### A. Meal preparation

The first set of experiments was based on a part of the CMU-MMAC database [41]. The CMU-MMAC database contains multimodal measures of human activity of subjects performing tasks involved in cooking and food preparation. Six synchronised cameras have been used to capture scenarios such as preparation of salad, pizza, eggs, and sandwich. Many types of tasks have been annotated within these scenarios. In our experiments, we considered the brownie preparation scenario. We have used twelve videos containing the full scenario, and sought to recognize 29 tasks described in Table I; the ground-truth annotations were taken from the dataset providers. Views from two cameras (7151020 and 7151062) were employed for that purpose (see Fig. 2).

To extract the spatiotemporal variations, we used pixel change history images to capture the motion history (see, e.g., [42]), and computed the complex Zernike moments $A_{00}, A_{11}, A_{20}, A_{22}, A_{31}, A_{33}, A_{40}, A_{42}, A_{44}, A_{51}, A_{53}, A_{55}, A_{60}, A_{62}, A_{64}, A_{66}$, for each of which we computed the norm and the angle. Additionally the center of gravity and the area of the found blobs were also used, making a total of 31 parameters, thus providing an acceptable scene reconstruction without a computationally prohibitive dimension. Zernike moments were calculated in rectangular regions of interest of approximately 15000 pixels in each image to limit the processing and allow real time feature extraction (performed at a rate of approximately 50-60 fps).

The employed HMMs comprised three states, each one having a single mixture component distribution, which facilitated fast algorithm execution with acceptable results. The streams were coupled using a Gaussian mixture of two components.

We randomly selected two full workflows for initial training (each containing 62 samples of all possible tasks), we used 2 different workflows to draw samples from (68 task samples in total) for the purposes of the active learning algorithm, and used the rest eight available workflows for testing (258 task samples in total). A graphical representation of the obtained success rates as new samples were included is given in Fig. 3.

Table I
Meal preparation tasks from the CMU-MMAC database, including their code and the total amount of samples in the twelve brownie preparation scenarios.

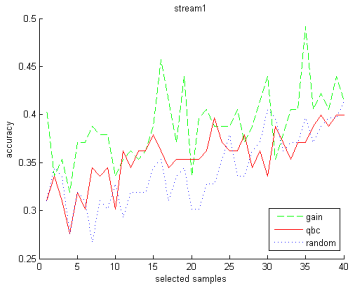
| task code | Task | total samples |
|---|---|---|
| 03 | close fridge | 11 |
| 06 | open brownie bag | 9 |
| 07 | open brownie box | 12 |
| 12 | open fridge | 11 |
| 14 | pour brownie bag into big bowl | 12 |
| 15 | pour oil into big bowl | 12 |
| 16 | pour oil into measuring cup small | 12 |
| 17 | pour water into big bowl | 12 |
| 18 | pour water into measuring cup big | 11 |
| 19 | put baking pan into oven | 12 |
| 24 | put pam into cupboard bottom right | 9 |
| 22 | put oil into cupboard bottom right | 10 |
| 27 | spray pam | 10 |
| 28 | stir big bowl | 12 |
| 30 | switch on | 12 |
| 31 | take baking pan | 12 |
| 32 | take big bowl | 12 |
| 33 | take brownie box | 12 |
| 34 | take egg | 11 |
| 35 | take fork | 12 |
| 37 | take measuring cup big | 12 |
| 38 | take measuring cup small | 12 |
| 39 | take oil | 10 |
| 40 | take pam | 9 |
| 42 | twist off cap | 11 |
| 43 | twist on cap | 12 |
| 44 | walk to counter | 11 |
| 45 | walk to fridge | 11 |
| 50 | crack egg on big bowl | 9 |

### B. Industrial part assembly

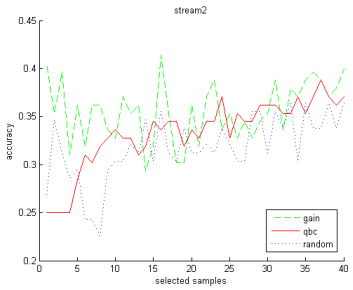
We used the WR dataset, and specifically the first two workflows pertaining to car assembly (see [43] for more details). The tasks to recognize in each of the workflows are the following:

1) Worker 1 picks up part 1 from rack 1 (upper) and places it on the welding cell; mean duration is 8-10 sec.
2) Worker 1 and worker 2 pick part 2a from rack 2 and place it on the welding cell.
3) Worker 1 and worker 2 pick part 2b from rack 3 and place it on the welding cell.
4) Worker 2 picks up spare parts 3a, 3b from rack 4 and places them on the welding cell.
5) Worker 2 picks up spare part 4 from rack 1 and places it on the welding cell.
6) Worker 1 and worker 2 pick up part 5 from rack 5 and place it on the welding cell.

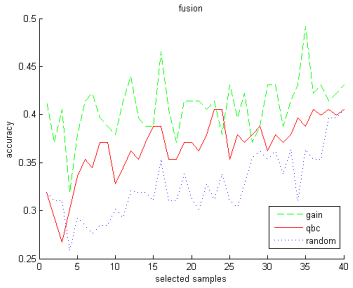
Each of the above tasks is a class that has to be recognized. The partial or total occlusions due to the racks make the task very difficult to complete with a single camera and therefore two views have been used (see Fig. 4), hence the need for a methodology allowing for the successful fusion of the information contained in tightly coupled times series.




(a) Accuracy of camera 7151020 steamwise model.




(b) Accuracy of camera 7151062 steamwise model.




(c) Accuracy fusing both cameras.

Figure 3. Success rates for the active learning methods compared to the random case, using a subset of the kitchen MMAC dataset. The x-axis is the number of selected samples for training, the y-axis is the respective accuracy on the test set.

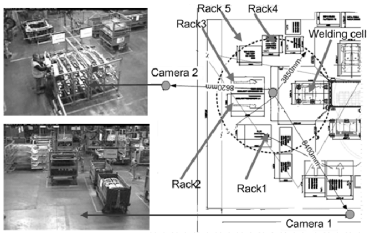
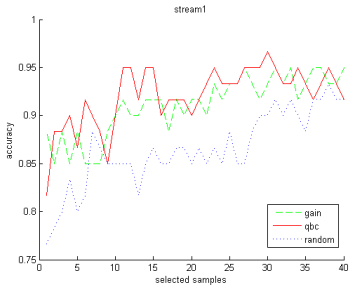


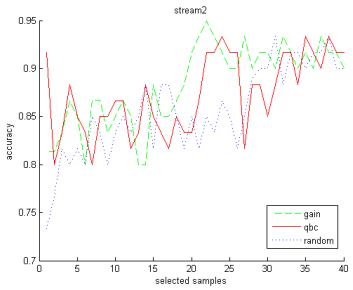

Figure 4. Schematic and camera views in the car assembly environment.

In our experiments, we have used two different workflows, each one comprising 20 sequences representing full assembly cycles and containing at least one of the considered behaviors. The total number of frames in each case was approximately 80000. Annotation of these frames has been performed manually. The second workflow is considered more difficult because the tasks may be executed in parallel, whereas in the first workflow the tasks were always executed sequentially. The same type of features were used as in the previous subsection. HMM configuration was similar to the previous experiment.

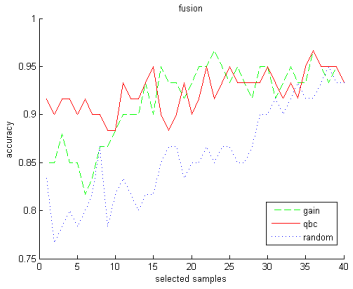
We randomly selected three full workflows for initial training (each containing all possible tasks), we used seven workflows to draw samples from (42 task samples in total) for the purposes of the active learning algorithm, and left the rest ten workflows for testing (60 task samples in total). The results for the first and second workflows are given in Fig. 5 and 6, respectively.




(a) Accuracy of camera 1 steamwise model.




(b) Accuracy of camera 2 steamwise model.

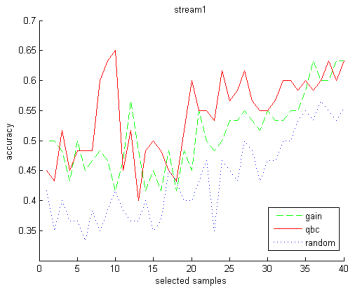



(c) Accuracy fusing both cameras.

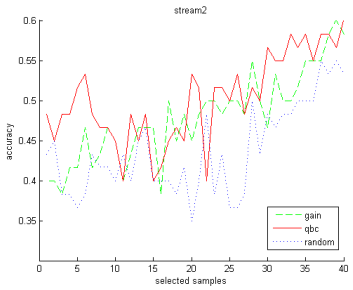
Figure 5. Success rates for the active learning methods compared to the random case, using the first workflow of the WR-dataset. The x-axis is the number of selected samples for training, the y-axis is the respective accuracy on the test set.

### C. Comparison to baseline classification methods

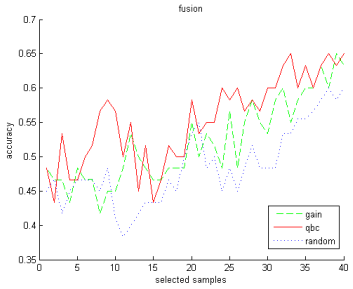
To verify the merit of the variational Bayesian approach towards observation fusion methods, we have included experimental comparisons of the variational Bayesian approach against the standard HMM and MFHMM models obtained using EM-based training. In our experiments, for both the first and second workflows of the WR dataset, the first ten scenarios were used for training and the rest scenarios for testing, while, in the case of the CMU-MMAC dataset, we trained with the first five scenarios and used the rest for testing. In all cases,




(a) Accuracy of camera 1 steamwise model.




(b) Accuracy of camera 2 steamwise model.




(c) Accuracy fusing both cameras.

Figure 6. Success rates for the active learning methods compared to the random case, using the second workflow of the WR-dataset. The x-axis is the number of selected samples for training, the y-axis is the respective accuracy on the test set.

three HMM states with a single component observation model were used for both the VB and EM methods. The results are displayed in Table II for models with Gaussian observation densities, and in Table III for models with Student's-t observation densities. As we observe, VB gives results which in most cases are much better than the EM algorithm for both the streamwise as well as the fused models. The higher accuracy comes, of course, at a higher computational cost. In all our experiments, classification using the VB models required between 4 and 5 times more time compared to the EM approach. Nevertheless, although higher, the computational time needed still remains of the same order of magnitude.

Furthermore, we have also compared to MCMC-based methods; for this purpose, we have considered the HMM model proposed in [44]. In our experiments, we used a truncation level of 10 states for this model, and imposed priors similar to our VB-based inference algorithm. Theoretically, higher accuracy is expected as the number of sampling iterations increases. Indeed, we observed this behavior in our experiments; nevertheless, to achieve similar or higher performance compared to the corresponding VB-based models, a very large number of iterations was needed, requiring too many computational resources, although the dimensionality of the problem was not too large. In Table II, we provide the results of the MCMC-based method of [44] for 10000 sampling iterations, which is a number of sampling iterations incurring reasonable computational costs (12 hours on an Intel Xeon 2.53GHz PC).

Finally, regarding the comparative computational costs of sequence classification using the proposed VB-based MFHMM model and simple streamwise models, we would like to mention that the costs of the proposed model are roughly equal to the sum of the costs of the corresponding streamwise models. Hence, in cases where two streams are used, our approach roughly imposes double the costs of a single streamwise model. This result was theoretically expected, considering that prediction in our model is conducted using Eq. (22).

### D. Discussion

In our experimental investigations, we evaluated the performance of the proposed information fusion scheme. Clearly, our fusion approach yielded improved results over methods using single-stream information. We also observed that the VB methods outperformed the respective EM-based ones for both the streamwise as well as the fused models. This result was theoretically expected since the latter models make point-estimates, which are more vulnerable to overfitting [27].

We also investigated the effectiveness of the proposed framework in an active learning setting. Two different active learning criteria were examined, namely information gain and query by committee. Using these methods, we were able to select the most appropriate samples to incorporate in model training. This process of sample selection was repeated until the maximum number of new samples was reached.

It has to be mentioned that QBC entails sampling of the model parameters, which may require a large number of experts. In our setting we used 30 experts, by drawing the same number of samples; increasing the number of experts would give more representative results, however the computational burden would increase proportionally. In our setting, the required execution time was almost the same for both methods for the selected amount of experts used from the QBC method.

Clearly, active learning outperformed random sample selection. To achieve the same performance, active learning methods require much less data than random selection. The differences in accuracy are bigger when adding only few samples. We have observed that both the gain and QBC criteria are able to select the samples that are closer to optimal in the sense of acquired information. As expected, we also observed that as more samples are labelled and added to the training set, the gap in performance compared to random selection tends to reduce. Furthermore, we noted that in most cases none of the proposed active learning methods could significantly outperform the other.

Table II

COMPARISON TO STANDARD EM APPROACHES USING THE GAUSSIAN OBSERVATION MODEL. COLUMNS EM-HMM1 AND EM-HMM2 PROVIDE THE ACCURACY OF THE EM-TRAINED STREAMWISE HMMS, AND EM-MFHMM PROVIDES THE ACCURACY OF THE EM-TRAINED MULTISTREAM FUSED HMM. THE CORRESPONDING RESULTS FOR MODELS TRAINED USING THE VARIATIONAL BAYESIAN APPROACH ARE PROVIDED IN COLUMNS VB-HMM1, VB-HMM2, AND VB-MFHMM, RESPECTIVELY. ACCURACY FOR MCMC-TRAINED STREAMWISE MODELS ARE PROVIDED IN MCMC-1 AND MCMC-2.

| Dataset | EM-HMM1 | EM-HMM2 | MFHMM | VB-HMM1 | VB-HMM2 | VB-MFHMM | MCMC-1 | MCMC-2 |
|---|---|---|---|---|---|---|---|---|
| CMU-MMAC | 39.49 | 35.90 | 41.03 | 43.08 | 37.95 | 44.62 | 42.13 | 29.23 |
| WR 1 | 90.00 | 70.00 | 90.00 | 95.00 | 86.67 | 96.67 | 78.00 | 71.00 |
| WR 2 | 55.71 | 37.14 | 63.33 | 63.33 | 56.67 | 68.33 | 35.00 | 45.00 |

Table III

COMPARISON TO STANDARD EM APPROACHES USING THE STUDENT'S-T OBSERVATION MODEL: COLUMNS EM-HMM1 AND EM-HMM2 PROVIDE THE ACCURACY OF THE EM-TRAINED STREAMWISE HMMS, AND EM-MFHMM PROVIDES THE ACCURACY OF THE EM-TRAINED MULTISTREAM FUSED HMM. THE CORRESPONDING RESULTS FOR MODELS TRAINED USING THE VARIATIONAL BAYESIAN APPROACH ARE PROVIDED IN COLUMNS VB-HMM1, VB-HMM2, AND VB-MFHMM, RESPECTIVELY.

| Dataset | EM-HMM1 | EM-HMM2 | MFHMM | VB-HMM1 | VB-HMM2 | VB-MFHMM |
|---|---|---|---|---|---|---|
| CMU-MMAC | 41.03 | 33.85 | 43.07 | 43.59 | 42.56 | 45.64 |
| WR 1 | 90.00 | 72.86 | 91.42 | 93.33 | 91.67 | 98.33 |
| WR 2 | 60.00 | 38.33 | 65.71 | 61.67 | 56.67 | 68.33 |

## V. CONCLUSIONS

In this work, we presented a novel variational Bayesian treatment of multistream fused hidden Markov models, with application to visual workflow recognition using multicamera networks. MFHMMs have been very successful in fusion of information from tightly interdependent data streams, with low computational requirements. In this work, we employed an elegant variational Bayesian treatment, which does not need large amounts of training data to guarantee dependable model estimation, since variational Bayes is much less prone to overfitting. Hence, despite the fact that the annotation of training data can be a major bottleneck, our VB-based method does not require large amount of them.

A major advantage of the proposed variational Bayesian treatment of MFHMMs over conventional approaches consists in the provision of a measure of confidence in the obtained model estimates. As we have shown, utilization of this information allows for the computationally efficient integration of the MFHMM into an active learning framework, by application of popular active learning criteria that would be either computationally cumbersome or even intractable were it not for the proposed variational Bayesian treatment.

## REFERENCES


[1] G. L. Foresti, C. Micheloni, L. Snidaro, P. Remagnino, and T. Ellis, "Active video-based surveillance systems," *IEEE Signal Proc. Magazine*, vol. 22, no. 2, pp. 25–37, 2005.

[2] G. L. Foresti, C. S. Regazzoni, and P. K. Varshney, *Multisensor Surveillance Systems: The Fusion Perspective*. Norwell, MA: Kluwer, 2003.

[3] N. Oliver, B. Rosario, and A. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831–843, 2000.

[4] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, pp. 141–151, 2000.

[5] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," *Proc. IEEE*, 1997.

[6] N. Oliver, E. Horvitz, and A. Garg, "Layered representations for learning and inferring office activity from multiple sensory channels," in *Proc. Int'l Conf. Multimodal Interfaces*, 2002.

[7] D. G. Stork and M. E. Hennecke, "Speech reading by humans and machines," in *NATO ASI Series F*. Springer Verlag, 1996, vol. 150.

[8] J. Triesch and C. von der Malsburg, "Democratic integration: Self-organized integration of adaptive cues," *Neural Comput.*, vol. 13, no. 9, pp. 2049–2074, 2001.

[9] O. Kahler, J. Denzler, and J. Triesch, "Hierarchical sensor data fusion by probabilistic cue integration for robust 3D object tracking," in *Proc. 6th IEEE Southwest Symp. Image Anal. and Interpret.*, 2004, pp. 216–220.

[10] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Comm.*, 2001.

[11] C. Vogler and D. Metaxas, "A framework for recognizing the simultaneous aspects of American sign language," *Computer Vision and Image Understanding*, vol. 81, no. 358-384, 2001.

[12] Z. Zeng, J. Tu, B. M. P. Jr., and T. S. Huang, "Audio–visual affective expression recognition through multistream fused HMM," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 570–577, 2008.

[13] K. Yamazaki and S. Watanabe, "Singularities in mixture models and upper bounds of stochastic complexity," *Neural Networks*, vol. 16, no. 7, pp. 1029–1038, 2003.

[14] C. Archambeau, J. Lee, and M. Verleysen, "On the convergence problems of the EM algorithm for finite Gaussian mixtures," in *Eleventh European symposium on artificial neural networks*, 2003, pp. 99–106.

[15] G. McLachlan and D. Peel, *Finite Mixture Models*. New York: Wiley Series in Probability and Statistics, 2000.

[16] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[17] J. Diebolt and C. Robert, "Estimation of finite mixture distributions through Bayesian sampling," *J. Roy. Statist. Soc. B*, vol. 56, pp. 363–375, 1994.

[18] S. Richardson and P. Green, "On Bayesian analysis of mixtures with unknown number of components," *J. Roy. Statist. Soc. B*, vol. 59, pp. 731–792, 1997.

[19] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*, M. Jordan, Ed. Dordrecht: Kluwer, 1998, pp. 105–162.

[20] C. Bishop and M. Tipping, "Variational relevance vector machines," in *Proc. 16th Conf. Uncertainty in Artificial Intelligence*, 2000, pp. 46–53.

[21] S. Roberts and W. Penny, "Variational Bayes for generalized autoregressive models," *IEEE Trans. Signal Processing*, vol. 50, pp. 2245–2257, 2002.

[22] V. Smidl and A. Quinn, "Mixture-based extension of the AR model and its recursive Bayesian identification," *IEEE Trans. Signal Processing*, vol. 53, pp. 3530–3542, 2005.

[23] C. Archambeau and M. Verleysen, "Robust Bayesian clustering," *Neural Networks*, vol. 20, pp. 129–138, 2007.

[24] M. Svensén and C. M. Bishop, "Robust Bayesian mixture modelling," *Neurocomputing*, vol. 64, pp. 235–252, 2005.

[25] Z. Ghahramani and M. Beal, "Variational inference for Bayesian mixture of factor analysers," *Advances Neural Information Processing Systems*, vol. 12, 1999.

[26] S. Chatzis, D. Kosmopoulos, and T. Varvarigou, "Signal modeling and classification using a robust latent space model based on $t$ distributions," *IEEE Trans. Signal Processing*, vol. 56, no. 3, March 2008.

[27] S. Chatzis and D. Kosmopoulos, "A variational Bayesian methodology for hidden Markov models utilizing Student's-t mixtures," *Pattern Recognition*, vol. 44, no. 2, pp. 295–306, 2011.

[28] I. Rezek and S. J. Roberts, "Ensemble hidden Markov models with extended observation densities for biosignal analysis," in *Probabilistic Modeling in Biomedicine and Medical Bioinformatics*, E. D. Husmeier, R. Dybowski, and S. Roberts, Eds. Springer Verlag, 2005.

[29] D. MacKay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, pp. 589–603, 1992.

[30] D. Cohn, Z. Ghahramani, and M. Jordan, "Active learning with statistical models," *J. Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.

[31] S. Chatzis, D. Kosmopoulos, and T. Varvarigou, "Robust sequential data modeling using an outlier tolerant hidden Markov model," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1657–1669, 2009.

[32] S. P. Luttrell, "The use of Bayesian and entropic methods in neural network theory," in *Maximum Entropy and Bayesian Methods*. Boston, MA: Kluwer, 1989, pp. 363–370.

[33] H. Pan, Z.-P. Liang, and T. S. Huang, "Estimation of the joint probability of multisensory signals," *Pattern Recogn. Lett.*, vol. 22, pp. 1431–1437, 2001.

[34] D. Povey and P. Woodland, "Minimum phone error and i-smoothing for improved discrimative training," in *Proc. ICASSP*, 2002.

[35] S. Raudys and A. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 3, pp. 252–264, 1991.

[36] M. A. Osborne, R. Garnett, and S. J. Roberts, "Active data selection for sensor networks with faults and changepoints," in *Proc. IEEE 24th International Conference on Advanced Information Networking and Applications (AINA 2010)*, 2010, pp. 533–540.

[37] S. Ji, B. Krishnapuram, and L. Carin, "Variational Bayes for continuous hidden Markov models and its application to active learning," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 522–532, 2006.

[38] D. MacKay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, pp. 589–603, 1992.

[39] Y. Freund, H. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Machine Learning*, vol. 28, pp. 133–168, 1997.

[40] H. Seung, M. Opper, and H. Smopolinsky, "Query by committee," in *Proc. Fifth Ann. ACM Workshop Computational Learning Theory*, 1992, pp. 287–294.

[41] F. D. la Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcada, and J. Macey., "Guide to the Carnegie Mellon University multimodal activity (CMU-MMAC) database," Carnegie Mellon University, Tech. Rep. CMU-RI-TR-08-22, July 2009.

[42] D. Kosmopoulos and S. Chatzis, "Robust visual behavior recognition," *Signal Processing Magazine, IEEE*, vol. 27, no. 5, pp. 34 –45, sept. 2010.

[43] A. Voulodimos, D. Kosmopoulos, G. Vasileiou, E. Sardis, A. Doulamis, V. Anagnostopoulos, C. Lalos, and T. Varvarigou, "A dataset for workflow recognition in industrial scenes," in *IEEE Int. Conference on Image Processing*, 2011, pp. 3310–3313.

[44] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "An HDP-HMM for systems with state persistence," in *Proc. International Conference on Machine Learning*, July 2008.